netarchive.dk

Newsletter – Year-1

Welcome to the first official newsletter from the netarchive.dk project. This newsletter covers the first year of the project (2005-2006). Our plan is to publish a newsletter every six months.

The netarchive.dk project started to harvest the Danish part of the internet on July 1st 2005 as a result of a revision of the Danish legal deposit law. The revision gave the national libraries in Denmark (the Royal Library in Copenhagen and the State and University Library in Aarhus) permission to collect and preserve Danish internet material.

The development of a distributed system to carry out this task began in the spring of 2004 and the system is still being further developed. This newsletter will reveal that we are now planning an open source release of the system.

The organisation of the project is structured so that staff at the two libraries undertakes the daily work with a daily manager coordinating all activities. Three departments at each library are involved: IT-services is responsible for running the infrastructure (servers and network), IT-development maintains the software and the National Collection departments (mainly librarians) initiate and monitor the harvests as well as carry out quality control on the harvested material.

The steering committee has the overall responsibility for netarchive.dk, and the Danish Ministry of Culture has put up an editorial board with members from the internet business field in Denmark to advise the project about collection strategies, etcetera.

Harvests in year-1

The first year has been exciting for all of us. As you may know our strategy is based on three different types of harvests:

1. Four annual snapshot harvests
2. Selective harvests of 80 websites with frequencies up to many times a day
3. Two to three annual event harvests of national events like elections or sports events

During our first year we have done three snapshot harvests, two event-based harvests – one concerning local elections in the autumn of 2005, and one covering the crisis that appeared as a result of the drawings of Mohammed. The selective harvests have collected material from approximately 40 websites - the number is going to increase to around 80 websites in the very near future.

One of our experiences with the Danish part of the internet is that 85% of all registered domains on the .dk TLD (around 700,000 domains) are smaller than 10Mbytes and only about 6,300 domains (less than 1%) are bigger than 500Mbytes. JPG pictures use the most disc space whereas - not surprisingly - PDF is the most used format for publishing documents. More details about our first snapshot harvest can be found in an article published on our website (http://netarchive.dk)

At the moment the harvesting is done by using four servers but as the amount of information on the internet increases, the number of harvester servers will have to meet this growing need.

We make use of the open source heritrix crawler which we have found to be a very powerful piece of software. Special needs for our setup have been developed by the Internet Archive on request. We have discovered that they are always willing to answer your questions and help you with your problems.

Bit Preservation

Our complete bit archive setup consists of servers running in both Copenhagen and Aarhus. The system automatically keeps two copies of everything we have harvested and regularly carries out automatic check-ups on all files in order to ensure that both copies are identical and have not been altered since the original was harvested. I'm very proud to announce that the archive is still 100% consistent even though the amount of data is already quite large and constantly growing. The active bit preservation system also includes mechanisms used for making corrections in either of the two archives if inconsistencies are discovered.

The archive contains 21,5 Tbytes of data from the three types of harvests:

- Snapshots: 16,374 GBytes
- Selective harvests: 2,214 GBytes
- Event harvests: 2,929 Gbytes

Technical Issues

Anyone who has ever tried to send a web crawler onto the internet knows that it can be a challenging task. We have had to find solutions to issues such as crawlertraps, web server overload due to hundred of thousands of items in shopping baskets, deletion of data on websites, etcetera. The problem with deletion of data turned out to reveal serious security issues on both websites so the fact that we harvest websites can actually be quite helpful to some of the websites we pay a visit.

The overall number of technical issues has been less than 40, so if you take into account that we have visited over 700,000 domains the number is actually quite small.

Naturally, like most new systems our system has had some initial problems, but it's becoming more and more stable due to continuous developments and improvements.

Recently we have incorporated the DeDuplicater module for heritrix developed by Iceland. Hopefully that will help us save up to 50% of the disc space we use today. With the ever-growing internet we are looking forward to seeing the result of this when the next snap shot harvest has been completed.

The netarchive.dk System in Open Source

During our first year we have been contacted by several institutions with enquiries about our system and the possible availability for them to use it for national obligations like ours. I'm therefore very proud to announce that we are now planning an open source release of our system.

At the moment we are in the process of adjusting the system so it is structured a little more modular and thus making it possible for us to release the software in different modules. The plan is to release the first module in the spring of 2007 and hopefully we can make the entire system available by the end of 2007.

Please visit our website – http://netarchive.dk - to stay up-dated on the process and download our code when it is available. On the website you will also find other articles related to the project and our activities. Please feel free to contact us by using netarkivet-svar@netarkivet.dk if you have any questions about the project.

Bjarne Andersen
Daily Manager